Some theoretical results on the grouped variables Lasso

Christophe Chesneau, Mohamed Hebiri

Some theoretical results on the Grouped Variables Lasso

Christophe Chesneau[1] and Mohamed Hebiri[2]

Abstract: We consider the linear regression model with Gaussian error. We estimate the unknown parameters by a procedure inspired from the Group Lasso estimator introduced in [21]. We show that this estimator satisfies a sparsity oracle inequality, i.e., a bound in terms of the number of non-zero components of the oracle vector. We prove that this bound is better, in some cases, than the one achieved by the Lasso and the Dantzig selector.

Key words and phrases: Lasso, Group Lasso, Variable selection, Sparsity, Oracle Inequality, Penalized least squares.

AMS 2000 Subject Classifications: Primary: 62J05, 62J07, Secondary: 62H20, 62F30.

1 Introduction

We consider the linear regression model

$$y_i = x_i\beta^* + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where the design $x_i = (x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^p$ is deterministic, $\beta^* = (\beta_1^*, \ldots, \beta_p^*)' \in \mathbb{R}^p$ is the unknown parameter vector of interest and $\varepsilon_1, \ldots, \varepsilon_n$, are i.i.d. centered Gaussian random variables with variance σ^2. We wish to estimate β^* in the sparse case, i.e. when many of its components are equal to zero. If we define the covariates $\xi_j = (x_{1,j}, \ldots, x_{n,j})'$, $j = 1, \ldots, p$, the sparsity of the model means that only a

[1] Université de Caen, LMNO, Campus II, Science 3, 14032, Caen, France
[2] Université Paris VII, LPMA, 175 rue du Chevaleret, 75013, Paris, France

small subset of $(\xi_j)_j$ is relevant for explaining the response y_i, $i = 1, \ldots, n$. We are mainly interested in the case where the number of the covariates p is much larger than the sample size n. In such a situation, the classical methods of estimation such as ordinary least squares are inconsistent. In the last decade, a wide variety of procedures has been developed for estimation and variable selection under sparsity assumption. Most popular procedures are of the form:

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \left\{ \|Y - X\beta\|_n^2 + \text{pen}(\beta) \right\}, \tag{1.2}$$

where $X = (x_1', \ldots, x_n')'$, $Y = (y_1, \ldots, y_n)'$, pen : $\mathbb{R}^p \to \mathbb{R}$ is a positive function measuring the complexity of the vector β and, for any vector $a = (a_1, \ldots, a_n)'$, $\|a\|_n^2 = n^{-1} \sum_{i=1}^n a_i^2$ (we denote by $\langle \cdot, \cdot \rangle_n$ the corresponding inner product in $\mathbb{R}^n$). When X is standardized, the Lasso procedure introduced in [18] is defined by (1.2) with $\text{pen}(\beta) = \lambda_{n,p} \sum_{i=1}^n |\beta_i|$, where $\lambda_{n,p}$ denotes a tuning parameter. This estimator is attractive as it performs both regression parameters estimation and variable selection. In the literature, the theoretical and empirical properties of the Lasso procedure have been extensively studied. See, for instance, [10], [17], [11], [13], [23] and [22], among others. Recent extensions of the Lasso and their performances can be found in [11], [16], [23], [24] and [19].

In this paper, we study a "grouped" version of the Lasso procedure. It is defined with a penalty of the form $\text{pen}(\beta) = \lambda_{n,p} \sum_{l=1}^L \sqrt{\sum_{j \in G_l} \|\xi_j\|_n^2 {\beta_j}^2}$, where the tuning parameter $\lambda_{n,p}$ depends on n and on p. It can be viewed as a slight modification of the Group Lasso procedure developed in [21]. For the sake of clarity, we call our modified Group Lasso: Grouped Variables Lasso. We measure its performance by considering a statistical approach derived from confidence balls. We aim to find the smallest bound $\varphi_{n,p}$ such that

$$\mathbb{P}\left(\|X\widehat{\beta} - X\beta^*\|_n^2 \leq C\, \varphi_{n,p} \right) \geq 1 - u_{n,p}, \tag{1.3}$$

where $\widehat{\beta}$ is the Grouped Variables Lasso estimator, $u_{n,p}$ is a positive sequence of the form $n^{-\alpha} p^{-\gamma}$ with $\alpha > 0$, $\gamma > 0$ and C is a positive constant which does not depend on n and p. The obtained rate $\varphi_{n,p}$ depends only on n, on p and on an index of sparsity of the model. From this point of view, the inequality (1.3) is a *Sparsity Oracle Inequality* (SOI) for the Grouped Variable Lasso estimator. Such

SOIs have already been investigated for other estimators ([5], [8], [14], [20] and [6]). As a benchmark, we use the SOIs provided for the Lasso estimator [3] and the Dantzig selector [2]. If we compare the corresponding $\varphi_{n,p}$, we remark that the one achieved by the Grouped Variables Lasso is smaller than the one achieved by the Lasso and the Dantzig selector. This illustrates the fact that, in some situations, the Grouped Variables Lasso exploits the sparsity of the model more efficiently than the Lasso and the Dantzig selector.

The rest of the paper is organized as follows. The Grouped Variables Lasso estimator is described in Section 2. Section 3 presents the assumptions made on the model. The theoretical performance of the considered estimator is investigated in Section 4. The proofs are postponed to Section 5.

2 The Grouped Variables Lasso (GVL) estimator

In this study, for any real number a, $[a]$ denotes the integer part of a. For convenience, we assume that $p/[\log p]$ is an integer. We define the Grouped Variables Lasso (GVL) estimator by

$$\widehat{\beta} = \underset{\beta\in\mathbb{R}^p}{\text{Argmin}} \left\{ \|Y - X\beta\|_n^2 + 2\sum_{l=1}^{L} \sqrt{\sum_{j\in G_l} w_{n,j}^2 {\beta_j}^2} \right\}, \tag{2.1}$$

where $L = p/[\log p]$, for any $j \in \{1, ..., p\}$,

$$w_{n,j} = \lambda_{n,p}\|\xi_j\|_n, \qquad \lambda_{n,p} = \kappa\sigma\sqrt{n^{-1}\log(np)}, \tag{2.2}$$

$\kappa \geq 2$ and, for any $l \in \{1, \ldots, L\}$,

$$G_l = \big\{k \in \{1, \ldots, p\} : (l-1)[\log p] + 1 \leq k \leq l[\log p]\big\}. \tag{2.3}$$

Note that $G = (G_l)_l$ is a partition of the set $\{1, \ldots, p\}$ such that, for any $l \in \{1, \ldots, L\}$, $\text{Card}(G_l) = [\log p]$. The GVL estimator is a slight modification of the Group Lasso estimator developed in [21]. The only differences are the choice of the blocks G_l and the fact that, in our setting, we do not assume that $X'_{G_l}X_{G_l} =$

$I_{\mathrm{Card}(G_l)}$, where X_{G_l} is the restriction of X on the block G_l. The length of each block, $\mathrm{Card}(G_l) = [\log p]$, is based on theoretical considerations. Further details are given in Section 4. Recent developments concerning the Group Lasso method can be found in [12] and [15].

For any real number a, we set $(a)_+ = \max(a, 0)$. If X is the identity matrix I_n (and, a fortiori, the model (1.1) is the standard Gaussian sequence model), each component of the GVL estimator $\widehat{\beta}$ in the block G_l can be expressed in the following form $\widehat{\beta}_i = \left(1 - \left(\kappa\sigma\sqrt{2n^{-1}\log n}\right) / \sqrt{\sum_{j\in G_l} {y_j}^2}\right)_+ y_i$. In this case, $\widehat{\beta}$ can be viewed as a slight modification of the blockwise Stein estimator. This construction enjoys powerful theoretical properties in various statistical approaches (oracle inequalities, (near) minimax optimality,...). See, for instance, [7].

3 Assumptions

Recall that $X = (x_{i,j})_{i,j}$ is the $n \times p$ design matrix and, for any $j \in \{1, \ldots, p\}$, $\xi_j = (x_{1,j}, \ldots, x_{n,j})'$. Let $\rho_p = (\rho_p(j,k))_{j,k}$ be the correlation matrix defined by

$$\rho_p(j,k) = \frac{\langle \xi_j, \xi_k \rangle_n}{\|\xi_j\|_n \|\xi_k\|_n}, \qquad (j,k) \in \{1, \ldots, p\}^2.$$

We now present three assumptions we need to establish a SOI for the GVL estimator. They relate to the correlation matrix ρ_p:

– *Assumption (A1). Consider the set* $\mathcal{S}_2^l = \{a = (a_j)_{j\in G_l} \in \mathbb{R}^{[\log p]};\ \sum_{j\in G_l} a_j^2 \ \leq\ 1\ \}$. *There exists a constant* $C_* \geq 1$ *independent of* n *and of* p *such that*

$$\max_{l=1,\ldots,L} \sup_{a\in\mathcal{S}_2^l} \left(\sum_{j\in G_l} \sum_{k\in G_l} a_j a_k \rho_p(j,k) \right) \leq C_*.$$

The second assumption must be satisfied for a subset $\mathcal{B} \subseteq \{1, \ldots, L\}$ to be specified later.

– *Assumption (A2)($\mathcal{B}$). The correlation matrix* ρ_p *satisfies*

$$\max_{l\in\mathcal{B}} \max_{m=1,\ldots,L} \sqrt{\sum_{j\in G_l} \sum_{\substack{k\in G_m \\ k\neq j}} \rho_p^2(j,k)} \leq (32)^{-1} \mathrm{Card}(\mathcal{B})^{-1}.$$

Remark 3.1 *The condition in Assumption (A1) is equivalent to say that the larger eigenvalue of the diagonal blocks of the matrix ρ_p (i.e. eigenvalues of the correlation matrices restricted to covariates in the same group) is bounded by C_*.*

Lemma 3.1 below determines a standard family of matrices satisfying Assumption (A1).

Lemma 3.1 *Let $X = (x_{i,j})_{i,j}$ be a $n \times p$ matrix and, for any $j \in \{1, \ldots, p\}$, $\xi_j = (x_{1,j}, \ldots, x_{n,j})'$. Suppose that, for any $j, k \in \{1, .., p\}$, we have*

$$\langle \xi_j, \xi_k \rangle_n = r_n z_j z_k b_{|j-k|},$$

where $r = (r_n)_n$ is a sequence of real numbers, $z = (z_u)_u$ denotes a positive sequence and $b = (b_u)_u$ denotes a sequence in $l_1(\mathbb{N})$ with $b_0 > 0$. Then X satisfies Assumption (A1) with $C_* = 1 + 2b_0^{-1}\|b\|_{l_1}$, where $\|b\|_{l_1} = \sum_{j=1}^{p} |b_j|$.

Here are some comments on Assumption (A2)($\mathcal{B}$). In our study, Assumption (A2)($\mathcal{B}$) only needs to be satisfied for a particular set $\mathcal{B} = \Theta_G \subseteq \{1, \ldots, L\}$ (to be defined in Subsection 4.1). This set characterizes the sparsity of the model. Note also that Assumption (A2)($\mathcal{B}$) can be viewed as an extension of the "local" mutual coherence condition considered by [4]. This "local" mutual coherence condition has been introduced by [9]. When we treat the case $p \geq n$, such coherence condition is standard as almost all SOIs provided in the literature need a similar condition.

Remark 3.2 *For any two sets $\mathcal{B}_1$ and $\mathcal{B}_2$ such that $\mathcal{B}_1 \subseteq \mathcal{B}_2 \subseteq \{1, \ldots, L\}$, Assumption (A2)($\mathcal{B}_2$) implies Assumption (A2)($\mathcal{B}_1$).*

Remark 3.3 *If $\mathcal{B} = \{1, \ldots, L\}$ then Assumption (A2)($\mathcal{B}$) implies Assumption (A1) with $C_* = 1 + (32)^{-1}$. This is a consequence of the Hölder inequality.*

An example of a $n \times p$ matrix $X = (x_{i,j})_{i,j}$ satisfying Assumptions (A1) and (A2)($\mathcal{B}$) for any $\mathcal{B} \subseteq \{1, \ldots, L\}$, is the one characterized by the equality $\langle \xi_j, \xi_k \rangle_n = n^{\nu} p^{-\alpha|j-k|}$, with $\nu \in \mathbb{R}$ and $\alpha \geq (\log(32)/\log p)+1$. Here is a concise proof: Thanks to Lemma 3.1, Assumption (A1) is satisfied for any constant $C_* \geq 1 + 2p/(p-1)$ (for instance, $C_* = 4$). Moreover, we have

$\max_{l=1,\ldots,L} \max_{m=1,\ldots,L} \sqrt{\sum_{j\in G_l} \sum_{\substack{k\in G_m \\ k\neq j}} p^{-2\alpha|j-k|}} \leq p^{-\alpha} \max_{l=1,\ldots,L} \mathrm{Card}(G_l) = p^{-\alpha+1}([\log p]/p) \leq (32)^{-1}L^{-1} \leq (32)^{-1}\,\mathrm{Card}(\mathcal{B})^{-1}$ and Assumption (A2)($\mathcal{B}$) is satisfied.

When $p \leq n$, Assumption (A2)($\mathcal{B}$) can replaced by the following:

– *Assumption (A3). Consider the $p \times p$ Gram matrix Ψ_n defined by $\Psi_n = \left(\langle \xi_j, \xi_k \rangle_n\right)_{j,k}$. For any $p \geq 2$, there exists a constant $c_p > 0$ such that the matrix Z defined by*

$$Z = \Psi_n - c_p\,\mathrm{diag}(\Psi_n),$$

is positive semi-definite.

Assumption (A3) is the same as in [4, Assumption (A3)]. Further details can be found in [4, Remarks 4-5]. Assumption (A3) is, for instance, always fulfilled for positive matrices Ψ_n. It is important to notice that this assumption can be helpful when the "group mutual coherence" assumption is not satisfied; Assumptions (A2)($\mathcal{B}$) and (A3) can recover different types of design matrices.

4 Theoretical properties

In this section, we investigate some theoretical properties of the GVL estimator. Notice that all the results include the case $p \geq n$.

4.1 Main results

Here we provide SOIs acheived by the GVL estimator. These SOIs take advantage of the group structure of the estimator. The key is the introduction of a *group sparsity set* Θ_G defined by:

$$\Theta_G = \left\{ l \in \{1, \ldots, L\} : \text{ there exists an integer } j_0 \in G_l \text{ such that } \beta^*_{j_0} \neq 0 \right\}, \quad (4.1)$$

where G_l is defined by (2.3). Such a set contains group indexes and characterizes the sparsity of the model. Indeed, the "sparser" the model is, the smaller the sparsity index $\mathrm{Card}(\Theta_G)$ is. Proposition 4.1 below provides an upper bound for the squared error of the GVL estimator. This bound brings into play the sparsity index inferred by the group sparsity set Θ_G.

Proposition 4.1 *We consider the linear regression model (1.1). Let $\Lambda_{n,p}$ be the random event defined by*

$$\Lambda_{n,p} = \left\{ \max_{l=1,\dots,L} \sqrt{\sum_{j \in G_l} w_{n,j}^{-2} V_j^2} \leq 2^{-1} \right\}, \tag{4.2}$$

where $V_j = n^{-1} \sum_{i=1}^n x_{i,j} \varepsilon_i$ and $w_{n,j}$ is defined by (2.2). Let $\widehat{\beta}$ be the GVL estimator defined by (2.1) and Θ_G be the group sparsity set defined by (4.1). Suppose that X satisfies Assumption $(A2)(\Theta_G)$. Then, on $\Lambda_{n,p}$, we have

$$\|X\widehat{\beta} - X\beta^*\|_n^2 \leq C\, n^{-1} \log(np)\, \mathrm{Card}(\Theta_G), \tag{4.3}$$

where $C = 16\kappa^2\sigma^2$.

The proof of Proposition 4.1 is based on the 'argmin' definition of the estimator $\widehat{\beta}$ and some technical inequalities. Theorem 4.1 below states that, under some assumptions on X, the SOI (4.3) is true with high probability.

Theorem 4.1 *We consider the linear regression model (1.1). Let $\widehat{\beta}$ be the GVL estimator defined by (2.1) and Θ_G be the group sparsity set defined by (4.1). Suppose that X satisfies Assumptions (A1) and (A2)(Θ_G). Then we have*

$$\mathbb{P}\left(\|X\widehat{\beta} - X\beta^*\|_n^2 \leq C\, n^{-1} \log(np)\, \mathrm{Card}(\Theta_G) \right) \geq 1 - u_{n,p}, \tag{4.4}$$

where $C = 16\kappa^2\sigma^2$ and $u_{n,p} = p(np)^{-(2^{-1}\kappa - 1)^2/(2C_*)}$, with C_* is the constant appearing in Assumption (A1).

The proof of Theorem 4.1 uses Proposition 4.1 and a concentration inequality of the form $\mathbb{P}(\Lambda_{n,p}^c) \leq u_{n,p}$, where $\Lambda_{n,p}^c$ denotes the complementary of the set (4.2).

Corollary 4.1 below states that, when $p \leq n$, Theorem 4.1 holds with Assumption (A3) instead of Assumption (A2)($\mathcal{B}$).

Corollary 4.1 *We consider the linear regression model (1.1). Let Θ_G be the group sparsity set defined by (4.1). Suppose that X satisfies Assumptions (A1) and (A3). Then the GVL estimator (2.1) satisfies the inequality (4.4) with $C = 16{c_p}^{-1}\kappa^2\sigma^2$, where c_p is the constant appearing in Assumption (A3).*

The proof of Corollary 4.1 is similar to the proof of Proposition 4.1.

4.2 Comparison with the Lasso and the Dantzig selector

A result similar to Theorem 4.1 has been proved for the Lasso estimator in [3], and for the Dantzig selector in [6]. Moreover [2] stated that the squared error of the Lasso and the Dantzig selector are equivalent up to a constant factor. In these works, the authors provided similar SOIs. The main difference lies in the sparsity index $\mathrm{Card}(\Theta_G)$. For both the Lasso estimator and the Dantzig selector, it is replaced by $\mathrm{Card}(\Theta^*)$, where $\Theta^* = \{j \in \{1, \dots, p\};\ \beta_j^* \neq 0\}$. Since

$$\mathrm{Card}(\Theta_G) \leq \mathrm{Card}(\Theta^*),$$

Theorem 4.1 states that, with high probability, the GVL estimator can have a smaller squared error than the Lasso estimator. This illustrates the fact that, in some cases, the GVL estimator exploits better the sparsity of the model than the Lasso estimator and the Dantzig selector. Moreover, $\mathrm{Card}(\Theta_G)$ can be asymptotically significatively smaller than $\mathrm{Card}(\Theta^*)$. For example, if $p = n$ and the unknown parameter vector $\beta^* = (\beta_1^*, \dots, \beta_n^*)'$ is defined by $\beta^* = (\underbrace{1, \dots, 1}_{\log n}, \underbrace{0, \dots, 0}_{n - \log n})$, then $\mathrm{Card}(\Theta_G) = 1$ whereas $\mathrm{Card}(\Theta^*) = \log n$.

5 Proofs

Proof of Lemma 3.1. For the sake of simplicity in exposition and without loss of generality, we work on the set $G_1 = \{1, \dots, [\log p]\}$. Let us notice that, for any $u \in G_1$, we have $\|\xi_u\|_n = z_u \sqrt{r_n b_0}$. Therefore, for any $(j,k) \in \{1, \dots, p\}^2$, we have $\rho_p(j,k) = b_0^{-1} b_{|j-k|}$. Hence

$$\begin{aligned}
&\sum_{j \in G_1} \sum_{k \in G_1} a_j a_k \rho_p(j,k) \\
&= b_0^{-1} \sum_{j=1}^{[\log p]} \sum_{k=1}^{[\log p]} a_j a_k b_{|j-k|} = \sum_{j=1}^{[\log p]} a_j^2 + 2 b_0^{-1} \sum_{j=2}^{[\log p]} \sum_{k=1}^{j-1} a_j a_k b_{j-k} \\
&\leq \sum_{j=1}^{[\log p]} a_j^2 + b_0^{-1} \sum_{j=2}^{[\log p]} \sum_{u=1}^{j-1} (a_j^2 + a_{j-u}^2) b_u.
\end{aligned}$$

For any $a \in \mathcal{S}_2^l$, we have $\sum_{j=1}^{[\log p]} a_j^2 \leq 1$. Therefore

$$\sum_{j=2}^{[\log p]} \sum_{u=1}^{j-1} a_j^2 b_u = \sum_{j=2}^{[\log p]} a_j^2 \sum_{u=1}^{j-1} b_u \leq \|b\|_{l_1}$$

and

$$\sum_{j=2}^{[\log p]} \sum_{u=1}^{j-1} a_{j-u}^2 b_u = \sum_{u=1}^{[\log p]-1} b_u \sum_{j=u+1}^{[\log p]} a_{j-u}^2 \leq \|b\|_{l_1}.$$

Hence

$$\sup_{a \in \mathcal{S}_2^l} \left(\sum_{j \in G_1} \sum_{k \in G_1} a_j a_k \rho_p(j,k) \right) \leq (1 + 2b_0^{-1}\|b\|_{l_1}) = C_*.$$

This inequality can easily be extended to any set G_l. Thus, the matrix X satisfies Assumption (A1) with $C_* = 1 + 2b_0^{-1}\|b\|_{l_1}$.

□

Proof of Proposition 4.1. By definition of the penalized estimator (2.1), for any $\beta \in \mathbb{R}^p$, we have

$$\begin{aligned} &\|X\widehat{\beta} - X\beta^*\|_n^2 + 2\sum_{l=1}^{L} \sqrt{\sum_{j \in G_l} w_{n,j}^2 \widehat{\beta}_j^2} - \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i x_i \widehat{\beta} \\ \leq \;& \|X\beta - X\beta^*\|_n^2 + 2\sum_{l=1}^{L} \sqrt{\sum_{j \in G_l} w_{n,j}^2 \beta_j^2} - \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i x_i \beta. \end{aligned}$$

Therefore, taking $\beta = \beta^*$, we obtain the following inequality:

$$\begin{aligned} \|X\widehat{\beta} - X\beta^*\|_n^2 \;\leq\; & 2\sum_{l=1}^{L} \left[\sqrt{\sum_{j \in G_l} w_{n,j}^2 (\beta_j^*)^2} - \sqrt{\sum_{j \in G_l} w_{n,j}^2 \widehat{\beta}_j^2} \right] \\ & + \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i x_i \left(\widehat{\beta} - \beta^* \right). \end{aligned} \tag{5.1}$$

Recall that $V_j = n^{-1}\sum_{i=1}^n x_{i,j}\varepsilon_i$ and using the Hölder inequality, we have on the event $\Lambda_{n,p}$

$$\begin{aligned}\frac{2}{n}\sum_{i=1}^n \varepsilon_i x_i\left(\widehat{\beta}-\beta^*\right) &= 2\sum_{l=1}^L\sum_{j\in G_l} V_j\left(\widehat{\beta}_j-\beta_j^*\right)\\ &\le 2\sum_{l=1}^L\sqrt{\sum_{j\in G_l} w_{n,j}^{-2}V_j^2}\sqrt{\sum_{j\in G_l} w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2}\\ &\le \sum_{l=1}^L\sqrt{\sum_{j\in G_l} w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2}. \end{aligned} \tag{5.2}$$

It follows from (5.1), (5.2) and the definition of the group sparsity set Θ_G (see (4.1)) that

$$\begin{aligned}&\|X\widehat{\beta}-X\beta^*\|_n^2+\sum_{l=1}^L\sqrt{\sum_{j\in G_l} w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2}\\ &\le 2\sum_{l=1}^L\sqrt{\sum_{j\in G_l} w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2}+2\sum_{l=1}^L\left[\sqrt{\sum_{j\in G_l} w_{n,j}^2(\beta_j^*)^2}-\sqrt{\sum_{j\in G_l} w_{n,j}^2\widehat{\beta}_j^2}\right]\\ &= 2\sum_{l\in\Theta_G}\sqrt{\sum_{j\in G_l} w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2}+2\sum_{l\in\Theta_G}\left[\sqrt{\sum_{j\in G_l} w_{n,j}^2(\beta_j^*)^2}-\sqrt{\sum_{j\in G_l} w_{n,j}^2\widehat{\beta}_j^2}\right].\end{aligned}$$

Therefore using the Minkowski inequality, we have

$$\begin{aligned}\|X\widehat{\beta}-X\beta^*\|_n^2+\sum_{l=1}^L\sqrt{\sum_{j\in G_l} w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2} &\le 4\sum_{l\in\Theta_G}\sqrt{\sum_{j\in G_l} w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2}\\ &\le 4\sqrt{\mathrm{Card}(\Theta_G)}\sqrt{\sum_{l\in\Theta_G}\sum_{j\in G_l} w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2}.\end{aligned} \tag{5.3}$$

Now, let us bound the term $\sum_{l\in\Theta_G}\sum_{j\in G_l} w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2$. By a simple decomposition, we have

$$\begin{aligned}\|X\widehat{\beta}-X\beta^*\|_n^2 = \sum_{l\in\Theta_G}\sum_{j\in G_l}\|\xi_j\|_n^2\left(\widehat{\beta}_j-\beta_j^*\right)^2+n^{-1}\sum_{i=1}^n\left(\sum_{l\notin\Theta_G}\sum_{j\in G_l} x_{i,j}(\widehat{\beta}_j-\beta_j^*)\right)^2\\ +R(\Theta_G),\end{aligned} \tag{5.4}$$

where

$$\begin{aligned} R(\Theta_G) &= 2\sum_{l\in\Theta_G}\sum_{m\notin\Theta_G}\sum_{j\in G_l}\sum_{k\in G_m}\langle\xi_j,\xi_k\rangle_n(\widehat{\beta}_j-\beta_j^*)(\widehat{\beta}_k-\beta_k^*) \\ &+ \sum_{l\in\Theta_G}\sum_{\substack{m\in\Theta_G\\ m\neq l}}\sum_{j\in G_l}\sum_{k\in G_m}\langle\xi_j,\xi_k\rangle_n(\widehat{\beta}_j-\beta_j^*)(\widehat{\beta}_k-\beta_k^*) \\ &+ \sum_{l\in\Theta_G}\sum_{j\in G_l}\sum_{\substack{k\in G_l\\ k\neq j}}\langle\xi_j,\xi_k\rangle_n(\widehat{\beta}_j-\beta_j^*)(\widehat{\beta}_k-\beta_k^*). \end{aligned}$$

Note that $R(\Theta_G)$ is such that

$$|R(\Theta_G)| \leq 2\sum_{l\in\Theta_G}\sum_{m=1}^{L}\sum_{j\in G_l}\sum_{\substack{k\in G_m\\ k\neq j}}|\langle\xi_j,\xi_k\rangle_n||\widehat{\beta}_j-\beta_j^*||\widehat{\beta}_k-\beta_k^*|.$$

Moreover, since $n^{-1}\sum_{i=1}^n\left(\sum_{l\notin\Theta_G}\sum_{j\in G_l}x_{i,j}(\widehat{\beta}_j-\beta_j^*)\right)^2\geq 0$, the equality (5.4) implies that:

$$\begin{aligned} &\sum_{l\in\Theta_G}\sum_{j\in G_l}w_{n,j}^2\left(\widehat{\beta}_j-\beta_j^*\right)^2 \\ &\leq \left(\kappa\sigma n^{-1}\sqrt{\log(np)}\right)^2 n\Big(\|X\widehat{\beta}-X\beta^*\|_n^2-R(\Theta_G)\Big) \\ &\leq \left(\kappa\sigma n^{-1}\sqrt{\log(np)}\right)^2 n\Big(\|X\widehat{\beta}-X\beta^*\|_n^2+|R(\Theta_G)|\Big) \\ &\leq \left(\kappa\sigma n^{-1}\sqrt{\log(np)}\right)^2 n\Big(\|X\widehat{\beta}-X\beta^*\|_n^2 \\ &\quad +2\sum_{l\in\Theta_G}\sum_{m=1}^{L}\sum_{j\in G_l}\sum_{\substack{k\in G_m\\ k\neq j}}|\langle\xi_j,\xi_k\rangle_n||\widehat{\beta}_j-\beta_j^*||\widehat{\beta}_k-\beta_k^*|\Big). \end{aligned} \tag{5.5}$$

Let us set $\Pi_{j,k} = w_{n,j}^{-1} w_{n,k}^{-1} \langle \xi_j, \xi_k \rangle_n$. The Cauchy-Schwarz inequality yields

$$
\begin{aligned}
&\sum_{l\in\Theta_G} \sum_{m=1}^{L} \sum_{j\in G_l} \sum_{\substack{k\in G_m \\ k\neq j}} |\langle \xi_j, \xi_k \rangle_n| \, |\widehat{\beta}_j - \beta_j^*| \, |\widehat{\beta}_k - \beta_k^*| \\
&= \sum_{l\in\Theta_G} \sum_{m=1}^{L} \sum_{j\in G_l} \sum_{\substack{k\in G_m \\ k\neq j}} |\Pi_{j,k}| \, w_{n,j} w_{n,k} \, |\widehat{\beta}_j - \beta_j^*| \, |\widehat{\beta}_k - \beta_k^*| \\
&\leq \sum_{l\in\Theta_G} \sum_{m=1}^{L} \sqrt{\sum_{j\in G_l} \sum_{\substack{k\in G_m \\ k\neq j}} \Pi_{j,k}^2} \sqrt{\sum_{j\in G_l} \sum_{k\in G_m} w_{n,j}^2 w_{n,k}^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2 \left(\widehat{\beta}_k - \beta_k^*\right)^2} \\
&\leq \sup_{l\in\Theta_G} \sup_{m=1,\ldots,L} \sqrt{\sum_{j\in G_l} \sum_{\substack{k\in G_m \\ k\neq j}} \Pi_{j,k}^2} \left(\sum_{l=1}^{L} \sqrt{\sum_{j\in G_l} w_{n,j}^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2} \right)^2 \\
&= B(\Theta_G).
\end{aligned}
$$

Combining (5.3), (5.5), the previous inequality and using an elementary inequality of convexity, we obtain

$$
\begin{aligned}
&\|X\widehat{\beta} - X\beta^*\|_n^2 + \sum_{l=1}^{L} \sqrt{\sum_{j\in G_l} w_{n,j}^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2} \\
&\quad\leq 4n^{1/2} \left(\kappa\sigma n^{-1}\sqrt{\log(np)}\right) \sqrt{\mathrm{Card}(\Theta_G)} \sqrt{\|X\widehat{\beta} - X\beta^*\|_n^2 + 2B(\Theta_G)} \\
&\quad\leq 4n^{1/2} \left(\kappa\sigma n^{-1}\sqrt{\log(np)}\right) \sqrt{\mathrm{Card}(\Theta_G)} \sqrt{\|X\widehat{\beta} - X\beta^*\|_n^2} \\
&\quad+ 4\sqrt{2}n^{1/2} \left(\kappa\sigma n^{-1}\sqrt{\log(np)}\right) \sqrt{\mathrm{Card}(\Theta_G) B(\Theta_G)}. \qquad (5.6)
\end{aligned}
$$

An application of Assumption (A2)($\mathcal{B}$), with $\mathcal{B} = \Theta_G$ yields

$$
4\sqrt{2}n^{1/2} \left(\kappa\sigma n^{-1}\sqrt{\log(np)}\right) \sqrt{\mathrm{Card}(\Theta_G) B(\Theta_G)} \leq \sum_{l=1}^{L} \sqrt{\sum_{j\in G_l} w_{n,j}^2 \left(\widehat{\beta}_j - \beta_j^*\right)^2}. \qquad (5.7)
$$

It follows from (5.6) and (5.7) that

$$
\|X\widehat{\beta} - X\beta^*\|_n^2 \leq 4n^{1/2} \left(\kappa\sigma n^{-1}\sqrt{\log(np)}\right) \sqrt{\mathrm{Card}(\Theta_G)} \|X\widehat{\beta} - X\beta^*\|_n.
$$

Therefore,

$$
\|X\widehat{\beta} - X\beta^*\|_n^2 \leq 16n \left(\kappa\sigma n^{-1}\sqrt{\log(np)}\right)^2 \mathrm{Card}(\Theta_G) = C\, n^{-1} \log(np)\, \mathrm{Card}(\Theta_G),
$$

where $C = 16\kappa^2\sigma^2$. This ends the proof of Proposition 4.1.

□

Proof of Theorem 4.1. We set $v_{n,j} = \sqrt{\sum_{i=1}^n x_{i,j}^2} = n^{1/2}\|\xi_j\|_n$. Thanks to Proposition 4.1, it is enough to prove that

$$\mathbb{P}\left(\max_{l=1,\dots,L}\sqrt{\sum_{j\in G_l} w_{n,j}^{-2}V_j^2} \geq 2^{-1}\right) \leq p(np)^{-(2^{-1}\kappa-1)^2/(2C_*)}.$$

We have

$$\begin{aligned}\mathbb{P}\left(\max_{l=1,\dots,L}\sqrt{\sum_{j\in G_l} w_{n,j}^{-2}V_j^2} \geq 2^{-1}\right) &\leq \sum_{l=1}^{L}\mathbb{P}\left(\sqrt{\sum_{j\in G_l} w_{n,j}^{-2}V_j^2} \geq 2^{-1}\right)\\ &\leq (p/[\log p])\max_{l=1,\dots,L}\mathbb{P}\left(\sqrt{\sum_{j\in G_l} v_{n,j}^{-2}V_j^2} \geq 2^{-1}\kappa\sigma n^{-1}\sqrt{\log(np)}\right). \quad (5.8)\end{aligned}$$

In order to bound this last term, we introduce the Borell inequality. For further details about this inequality, see, for instance, [1].

Lemma 5.1 (The Borell inequality) *Let $\mathcal{D}$ be a subset of $\mathbb{R}$ and $(\eta_t)_{t\in\mathcal{D}}$ be a centered Gaussian process. Suppose that*

$$\mathbb{E}\left(\sup_{t\in\mathcal{D}}\eta_t\right) \leq N \quad \textit{and} \quad \sup_{t\in\mathcal{D}} Var(\eta_t) \leq Q.$$

Then, for any $x > 0$, we have

$$\mathbb{P}\left(\sup_{t\in\mathcal{D}}\eta_t \geq x + N\right) \leq \exp(-x^2/(2Q)).$$

Let us consider the set $\mathcal{S}_2^l$ defined by $\mathcal{S}_2^l = \{a = (a_j)_{j\in G_l} \in \mathbb{R}^{[\log p]};\ \sum_{j\in G_l} a_j^2 \leq 1\}$, and the centered Gaussian process $\mathcal{Z}(a)$ defined by

$$\mathcal{Z}(a) = \sum_{j\in G_l} a_j V_j v_{n,j}^{-1}.$$

By an argument of duality, we have

$$\sup_{a\in\mathcal{S}_2^l}\mathcal{Z}(a) = \sup_{a\in\mathcal{S}_2^l}\sum_{j\in G_l} a_j v_{n,j}^{-1}V_j = \sqrt{\sum_{j\in G_l} v_{n,j}^{-2}V_j^2}.$$

In order to use Lemma 5.1, let us investigate the upper bounds for $\mathbb{E}(\sup_{a\in\mathcal{S}_2^l}\mathcal{Z}(a))$ and $\sup_{a\in\mathcal{S}_2^l} Var(\mathcal{Z}(a))$, in turn.

The upper bound for $\mathbb{E}(\sup_{a\in\mathcal{S}_2^l}\mathcal{Z}(a))$. Since $V_j \sim \mathcal{N}\left(0, \sigma^2 n^{-1}\|\xi_j\|_n^2\right)$, the Cauchy-Schwarz inequality yields

$$\begin{aligned}\mathbb{E}(\sup_{a\in\mathcal{S}_2^l}\mathcal{Z}(a)) &= \mathbb{E}\left(\sqrt{\sum_{j\in G_l} v_{n,j}^{-2}V_j^2}\right) \leq \sqrt{\sum_{j\in G_l} v_{n,j}^{-2}\mathbb{E}(V_j^2)} \\ &= \sqrt{\sum_{j\in G_l} v_{n,j}^{-2}(\sigma^2 n^{-1}\|\xi_j\|_n^2)} = \sigma n^{-1}\sqrt{\log p}.\end{aligned}$$

So, we set $N = \sigma n^{-1}\sqrt{\log p}$.

The upper bound for $\sup_{a\in\mathcal{S}_2^l} Var(\mathcal{Z}(a))$. We have

$$Var(\mathcal{Z}(a)) = \sum_{j\in G_l}\sum_{k\in G_l} a_j a_k v_{n,j}^{-1} v_{n,k}^{-1}\mathbb{E}(V_j V_k),$$

with $\mathbb{E}(V_jV_k) = n^{-2}\sum_{u=1}^n\sum_{v=1}^n x_{u,j}x_{v,k}\mathbb{E}(\epsilon_u\epsilon_v) = \sigma^2 n^{-1}\langle\xi_j,\xi_k\rangle_n$. This with Assumption (A1) imply

$$\sup_{a\in\mathcal{S}_2^l} Var(\mathcal{Z}(a)) = \sigma^2 n^{-2}\sup_{a\in\mathcal{S}_2^l}\left(\sum_{j\in G_l}\sum_{k\in G_l} a_j a_k \rho_p(j,k)\right) \leq C_*\sigma^2 n^{-2}.$$

So, we set $Q = C_*\sigma^2 n^{-2}$.

Combining the obtained values of N and Q with Lemma 5.1, for any $l \in \{1,\ldots,L\}$, we have

$$\begin{aligned}&\mathbb{P}\left(\sqrt{\sum_{j\in G_l} v_{n,j}^{-2}V_j^2} \geq 2^{-1}\kappa\sigma n^{-1}\sqrt{\log(np)}\right) \\ &\leq \mathbb{P}\left(\sqrt{\sum_{j\in G_l} v_{n,j}^{-2}V_j^2} \geq (2^{-1}\kappa-1)\sigma n^{-1}\sqrt{\log(np)} + \sigma n^{-1}\sqrt{\log p}\right) \\ &= \mathbb{P}\left(\sup_{t\in\mathcal{D}}\eta_t \geq (2^{-1}\kappa-1)\sigma n^{-1}\sqrt{\log(np)} + N\right) \\ &\leq \exp\left(-(2^{-1}\kappa-1)^2\sigma^2 n^{-2}\log(np)/(2Q)\right) = (np)^{-(2^{-1}\kappa-1)^2/(2C_*)}. \quad (5.9)\end{aligned}$$

Putting (5.8) and (5.9) together, we obtain

$$\mathbb{P}\left(\max_{l=1,\dots,L}\sqrt{\sum_{j\in G_l} w_{n,j}^{-2}V_j^2}\geq 2^{-1}\right)\leq p(np)^{-(2^{-1}\kappa-1)^2/(2C_*)}=u_{n,p}.$$

This ends the proof of Theorem 4.1.

□

Acknowledgement. We would like to thank Professor Alexander Tsybakov and Professor Nicolas Vayatis for insightful comments.

References

[1] Adler, R. J. *An introduction to continuity, extrema, and related topics for general Gaussian processes.* Institute of Mathematical Statistics Lecture Notes—Monograph Series, 12. Institute of Mathematical Statistics, Hayward, CA, 1990.

[2] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. Simultaneous analysis of lasso and dantzig selector. *preprint* (2007).

[3] Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. Aggregation and sparsity via l_1 penalized least squares. In *Learning theory*, vol. 4005 of *Lecture Notes in Comput. Sci.* Springer, Berlin, 2006, pp. 379–391.

[4] Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. Aggregation for Gaussian regression. *Ann. Statist. 35*, 4 (2007), 1674–1697.

[5] Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat. 1* (2007), 169–194.

[6] Candes, E. J., and Tao, T. The dantzig selector: statistical estimation when p is much larger than n. *To appear in The Annals of Statistics.* (2007).

[7] Cavalier, L., and Tsybakov, A. B. Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Methods Statist. 10*, 3 (2001), 247–282.

[8] Dalalyan, A., and Tsybakov, A. B. Aggregation by exponential weighting and sharp oracle inequalities. *20th Annual Conference on Learning Theory, COLT 2007 Proceedings. Lecture Notes in Computer Science 4539 Springer 2007* (2007), 97–111.

[9] Donoho, D. L., Elad, M., and Temlyakov, V. N. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory 52*, 1 (2006), 6–18.

[10] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *Ann. Statist. 32*, 2 (2004), 407–499. With discussion, and a rejoinder by the authors.

[11] Fan, J., and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc. 96*, 456 (2001), 1348–1360.

[12] Kim, Y., Kim, J., and Kim, Y. Blockwise sparse regression. *Statist. Sinica 16*, 2 (2006), 375–390.

[13] Knight, K., and Fu, W. Asymptotics for lasso-type estimators. *The Annals of Statistics 28*, 5 (2000), 1356–1378.

[14] Koltchinskii, V. Sparsity in penalized empirical risk minimization. *Preprint*.

[15] Meier, L., van de Geer, S., and Bühlmann, P. The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B. 70*, 1 (2008), 53–71.

[16] Meinshausen, N. Lasso with relaxation. *Technical Report* (2005).

[17] Meinshausen, N., and Bühlmann, P. High dimensional graphs and variable selection with the lasso. *Ann. Statist. 34*, 3 (2006), 1436–1462.

[18] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58*, 1 (1996), 267–288.

[19] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol. 67*, 1 (2005), 91–108.

[20] Van de Geer, S. High-dimensional generalized linear models and the lasso. *To appear in The Ann. Statist.* (2007).

[21] Yuan, M., and Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol. 68*, 1 (2006), 49–67.

[22] Zhao, P., and Yu, B. On model selection consistency of Lasso. *J. Mach. Learn. Res. 7* (2006), 2541–2563.

[23] Zou, H. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc. 101*, 476 (2006), 1418–1429.

[24] Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol. 67*, 2 (2005), 301–320.